Preventive Veterinary Medicine 89 (2009) 25–33

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Population inferences from targeted sampling with uncertain epidemiologic information

Michael S. Williams [a,*], Eric D. Ebel [a,1], Scott J. Wells [b]

[a] Risk Assessment and Residue Division, Office of Public Health Science, Food Safety Inspection Service-USDA, 2150 Centre Avenue, Building D, Fort Collins, CO 80526, USA
[b] Department of Veterinary Population Medicine, University of Minnesota, 1365 Gortner Avenue, Saint Paul, MN 55108, USA

ARTICLE INFO

Article history:
Received 14 November 2007
Received in revised form 24 October 2008
Accepted 30 December 2008

Keywords:
Model-based inference
Prevalence
Disease detection
Surveillance

ABSTRACT

Targeted sampling is an increasingly popular method of data collection in animal-based epidemiologic studies. This sampling approach allows the user to exclusively choose samples from subpopulations that have a higher likelihood of the disease of interest. This is achieved by selecting animals from a subpopulation that exhibits some characteristic that indicates a higher probability of the presence of the disease. Inferences drawn from a targeted sample require information regarding the epidemiology of the disease under surveillance, which is generally not known with certainty. This study describes estimators for both the detection of disease and the estimation of prevalence when targeted sampling is employed. Modifications of these estimators are provided that account for the uncertainty in the parameters that describe the epidemiology of the disease. Results of a simulation study are provided to illustrate the effect of the uncertainty in these parameters.

Published by Elsevier B.V.

1. Introduction

Targeted sampling is an increasingly popular approach for surveillance applications where the prevalence of a disease is low (Christensen and Gardner, 2000; OIE, 2006; Tavornpanich et al., 2006; Prattley et al., 2007a,b). An advantage of targeted sampling is that samples can be collected from a small number of subpopulations while ignoring other subpopulations altogether. These subpopulations are defined by observable characteristics that indicate a different probability of the presence of disease. Targeted sampling has become a topic of considerable interest because its use often results in a substantial reduction in the sample size required to detect a disease with known confidence when the prevalence exceeds some predetermined threshold.

Two key aspects differentiate targeted sampling from other common sampling approaches, such as stratified sampling and multi-stage or multi-phase sampling designs (e.g., Cochran, 1977; Cameron and Baldock, 1998); (1) samples can be drawn exclusively from the targeted subpopulation(s) rather than all subpopulations, and (2) the inferences drawn from a sample rely on knowledge about the epidemiology of the disease, rather than auxiliary information that would traditionally be known as part of the sampling frame or estimated by a sample drawn from the same population.

Targeted sampling applications assign a *point value* to each animal that is sampled (Cannon, 2002). The point value for each animal is based on epidemiologic characteristics. An interpretation of the point value is that the number of points assigned to an animal from a targeted subpopulation represents the number of animals, randomly selected from the entire population, in order to achieve an equivalent inference.

This study shows that inferences from a targeted sample require knowledge of two epidemiologic parameters; the risk ratio associated with the characteristic

* Corresponding author. Tel.: +1 970 492 7189.
E-mail addresses: mike.williams@fsis.usda.gov (M.S. Williams), eric.ebel@fsis.usda.gov (E.D. Ebel), Wells023@umn.edu (S.J. Wells).
[1] Tel.: +1 970 492 7187.

0167-5877/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.prevetmed.2008.12.008

and the proportion of the population that exhibits the characteristic. Nevertheless, these parameters are generally unknown values and cannot be directly estimated for the population because the disease either does not exist in the population or it exists at such a low level that collection of sufficient data would be impractical. Instead, these values are often determined from studies performed on similar populations or the uncertainty is characterized by expert opinion. Regardless of the source of information, inferences drawn from a sample must account for the uncertainty in these epidemiologic parameters.

The goal of this study is to provide a theoretical and practical description of targeted sampling. Disease surveillance systems are designed to either detect disease within a population (conditioned on its existence at some predetermined level) or to estimate disease prevalence within the population. During the surveillance design phase, determination of the sample size to achieve these objectives follows necessarily different, but readily available, algorithms (Cannon, 2002). Typically sampling to detect disease requires a smaller number of samples than sampling to estimate prevalence with a reasonable level of precision. The sample size algorithms require some assumption, *a priori*, about the occurrence of disease in the population. Despite the initial differences during surveillance design, however, the objectives of the two approaches can often switch during the data analysis phase. Once sample results are available, a sample that was intended to detect disease can be used to estimate prevalence in the population (particularly if one or more samples are found to be from diseased subjects); furthermore, if a sample was designed to estimate prevalence but finds no diseased subjects, the results can be used to determine the confidence in detecting disease for a putative prevalence level. Therefore, application of targeted sampling must consider the design and analysis of surveillance systems for both of these objectives. Because many surveillance programs are primarily established to detect disease, the applications in this study will assume that sample size is initially determined based on sampling to detect algorithms. Regardless of the size of the sample or purpose of sampling, however, methodology for estimating prevalence from targeted sampling is also needed.

In this study methods are developed to determine the point value to assign to each animal, as well as the necessary sample size for detection applications. Estimators of population disease prevalence are evaluated. These topics are first addressed when perfect information regarding the epidemiology of the disease is available, then modifications that account for uncertainty in the epidemiologic parameters are provided. Finally, a simulation study illustrates some of the potential advantages and pitfalls of the targeted sampling approach compared to simple random sampling.

2. Outline of the targeted sampling approach

2.1. Background information and definition of epidemiologic parameters

Assume there is a large population of N animals. For example, N might be the 10 million sheep in the US. There is interest in demonstrating, with a high degree of confidence, that the prevalence of a disease in the population is less than some predetermined level. This prevalence, P, is the *design prevalence*, which is a user-defined threshold that plays a key role in determining the sample size required to detect the disease. To demonstrate that the prevalence of the disease in a population is less than P requires determining the number of samples, n, so that if all of the samples are negative, one can conclude that the level of disease is below the predetermined level with a pre-specified level of confidence, which is usually $1-\alpha = 0.95$ or 0.99 (i.e., if the prevalence of the disease is P, the probability that one or more diseased animals will be sampled is $1-\alpha$). This is done using the following logic. The probability of sampling a healthy animal at random is $p(\text{the animal is healthy}) = (1-P)$. Given that $N \gg n$, this result can be extended to a random sample of n animals so that $p(\text{all sampled animals are healthy}) \approx (1-P)^n$.

Then

$$\begin{aligned} &p(1 \text{ or more diseased animals are sampled}) \\ &\quad = 1 - p(\text{all sampled animals are healthy}) \\ &\quad \approx 1 - (1-P)^n. \end{aligned}$$

Setting

$$1-\alpha = 1-(1-P)^n$$

and solving for n gives

$$n = \frac{\ln \alpha}{\ln(1-P)}.$$

This provides the necessary sample size to conclude that the level of disease is less than P with confidence level $1-\alpha$.

The fundamental idea of targeted sampling is that it is possible to sample a subpopulation with an increased prevalence and reduce the required sample size. Suppose an objective characteristic exists for the disease.[2] Let T be the symbol for the characteristic and animals with the characteristic are more likely to have the disease. An example of a characteristic could be the face color of sheep for predicting the presence of scrapie (NAHMS, 2003 found that black-faced sheep have a much higher probability of scrapie infection than white-faced sheep). Those individuals not possessing the characteristic are denoted by O. For the purpose of developing the methodology, assume the prevalence of the disease in the general population and the design prevalence are equal. In this example, there are just two subpopulations of interest (i.e., those individuals with and without the characteristic) and the prevalence levels in the two subpopulations are P_T and P_O, which will be referred to as the prevalence within the targeted and non-targeted subpopulations, respectively. The relationship between these prevalence levels is $P_T > P > P_O$, so the probability that a sampled individual has the disease depends on the subpopulation from which it was selected. If a random sample of animals is drawn from the targeted

[2] Disease is arbitrarily chosen as the condition of interest. This development could equally apply to infection, carrier status or some other case definition.

subpopulation there will exist an integer-valued sample size $n_T < n$ such that

$$1 - \alpha = 1 - (1 - P)^n \cong 1 - (1 - P_T)^{n_T} \quad (1)$$

Making inferences from a targeted sample requires knowledge of epidemiologic parameters that describe the disease. These parameters are used to relate the prevalence in the general and targeted populations. The first parameter is the risk ratio associated with T and is defined as

$$RR = \frac{P(diseased|T)}{P(diseased|O)} = \frac{P_T}{P_O} \quad (2)$$

where $RR > 1$ for the targeted subpopulation. It is assumed that this value is objectively determined (e.g., research studies, a population survey, or expert opinion). The second parameter is the proportion of the population with the characteristic, denoted by f_T. The prevalence in the subpopulations is related to the population prevalence by the weighted average $P = f_T P_T + (1 - f_T)P_O$.

2.2. A points concept for targeted sampling

The concept of using a points-based system for demonstrating disease freedom is an intuitive approach for targeted sampling (Cannon, 2002). This approach assigns a relative value to each sample. Sample elements drawn from a subpopulation that has a higher prevalence receive a higher point value than randomly selected animals drawn from the general population or from a non-targeted population (i.e., samples that are more beneficial in the search for disease are more valuable). The number of points ascribed to an individual element equals the relative value of that sample element in comparison to a randomly sampled element from the general population. For example, if a sample size of $n = 300$ was required to detect one or more diseased individuals in a random sample from the general population, but a targeted sample only requires $n_T = 100$ animals drawn from the targeted subpopulation to detect one or more diseased individuals, then the number of points (γ) ascribed to each animal in the targeted sample is 3. Using this approach, the point value assigned to a targeted animal is such that

$$n = \gamma n_T.$$

Unlike a traditional sample design, which requires that every animal in the population has a nonzero probability of being selected, targeted sampling can be designed to focus solely on the targeted subpopulation(s). To make inferences about the general population, however, the point value must link inferences drawn from the targeted subpopulation(s) back to the general population.

If the goal is to estimate the prevalence of the disease in the general population from a targeted sample drawn from only a single subpopulation, then a value for γ must be defined such that the estimator for prevalence, based on a sample drawn from the targeted subpopulation, will be equivalent to a simple random sample (SRS) from the general population. Following this logic

$$\hat{P} = \frac{\text{number of diseased animals in a SRS sample}}{n} = \frac{\text{number of diseased animals in a targeted sample}}{\gamma n_T} = \frac{\hat{P}_T}{\gamma}.$$

This expression leads to the conclusion that the number of points to ascribe to a targeted animal relates the prevalence in the targeted subpopulation to the prevalence in the general population, with

$$\gamma = \frac{P_T}{P}.$$

The same logic for deriving points can also be applied to the problem of disease detection. In this application, the definition of γ is equivalent to solving the formula

$$1 - (1 - P)^n = 1 - (1 - P_T)^{n_T} = 1 - (1 - P_T)^{n/\gamma}.$$

This formula leads to the solution $\gamma = \ln(1 - P_T)/\ln(1 - P)$, with

$$\frac{\ln(1 - P_T)}{\ln(1 - P)} \neq \frac{P_T}{P}.$$

Prattley et al. (2007a,b) provide a different formulation for γ, while a fourth relationship can be derived if the goal is to design a survey where the precision of the estimated prevalence from a simple random sample matches that of a targeted sample.

At first glance this is a troublesome conclusion until one realizes that all of the interpretations of γ are approximately equal for low-prevalence surveillance applications because

$$\gamma = \frac{\ln(1 - P_T)}{\ln(1 - P)} \approx \frac{P_T}{P},$$

whenever both P_T and P are close to 0. The ratio of these two formulations of γ over a wide range of P_T and P values is given in Fig. 1, where substantial differences are apparent. However, the agreement between the two definitions of γ is good (i.e., the ratio is close to one)

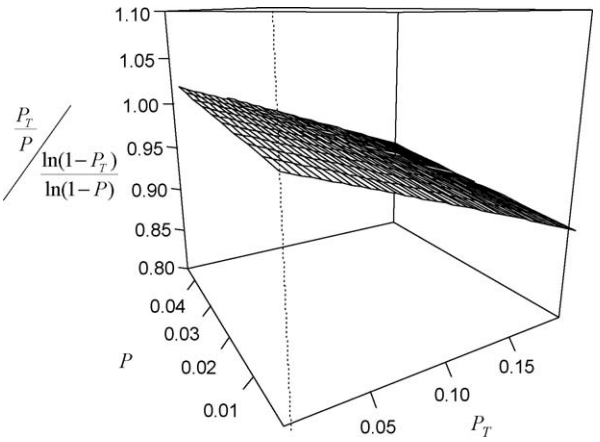

Fig. 1. Agreement between the point values (γ) derived from an application of sampling for detection versus the point values necessary for unbiased prevalence estimation. The range of possible prevalence levels ranges from nearly 0 to 0.5.

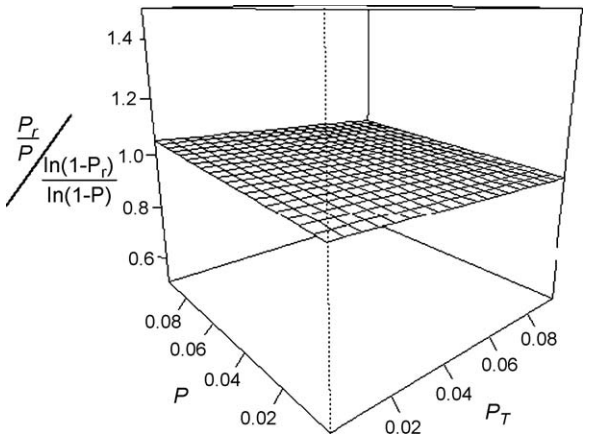

Fig. 2. Agreement between the point values (γ) for a sampling for detection application and the point values for prevalence estimation when all possible prevalence levels are restricted to being less than 0.1. The flat surface indicates close agreement in point values, regardless of the application.

when the range of P_T and P values are limited to being less than 0.1 (Fig. 2).

For the remainder of this study, the point value per sample will be $\gamma = P_T/P$. The motivation for using this formula is that points per sample can be explicitly related to epidemiologic parameters that describe the disease. Specifically, the number of points to ascribe to each targeted animal is given by,

$$\gamma = \frac{P_T}{P} = \frac{P_T/P_O}{(f_T P_T + f_O P_O)/P_O} = \frac{RR}{f_T RR + (1 - f_T)}.$$

A similar formula applies if multiple targeted subpopulations were considered. Suppose characteristics exist to divide the population into $L-1$ subpopulations, with the Lth subpopulation comprising those animals that did not exhibit any of the characteristics of interest (i.e., apparently healthy animals). In this case f_{T_l} and RR_{T_l} are required for each $l = 1,\ldots L-1$ targeted subpopulations and the points per targeted sample are given by

$$\gamma_l = \frac{P_{T_l}}{P} = \frac{RR_{T_l}238}{f_{T_l} RR_{T_l} + (1 - f_{T_l})}$$

for $l = 1,\ldots L-1$. Furthermore, if multiple targeted subpopulations are available for sampling, then the optimal allocation of sampling resources is to preferentially select animals from the subpopulation(s) with the highest point value. If the cost of sampling increases with accumulated numbers of samples (e.g., the cost of locating particular targeted samples becomes more difficult after collecting some number of samples), however, it may be more cost-effective to select samples from targeted subpopulations with lower values of γ_l.

2.3. Points and sampling from the non-targeted subpopulation

Given the prevalence in the general population (P), the effect of dividing the population into the targeted and non-targeted subpopulations is to increase the prevalence in the targeted subpopulation and decrease the prevalence in the non-targeted subpopulation. This implies that drawing a sample from the non-targeted subpopulation is less effective for finding disease than drawing a sample of equal size from the general population. The point value attributed to each sampled animal from the non-targeted portion of the population is

$$\gamma_O = \frac{P_O}{P} = \frac{1}{f_T RR + (1 - f_T)}.$$

The point value assigned to each sample from the non-targeted population is less than 1 given that RR is the risk ratio for the targeted population given in Eq. (2).

The number of points ascribed to animals that are apparently healthy, regardless of the number of different subpopulation, is given by

$$\gamma_O = \frac{1 - \sum_{l=1}^{L-1} f_{T_l}\gamma_l}{1 - \sum_{l=1}^{L-1} f_{T_l}},$$

where L denotes the total number of subpopulations and $l = L$ is the subpopulation of animals that have do not exhibit any of the characteristics of interest (i.e., the subpopulation of apparently healthy animals). Note that this equation reflects the constraint that

$$P = \sum_{l=1}^{L} f_l P_l.$$

2.4. Combining negative surveillance results from multiple subpopulations

A framework for combining results is necessary when samples are collected from multiple subpopulations. The term surveillance sensitivity is sometimes used to describe the probability of detecting disease in a population. Eq. (1) is a special case of the more general formula for determining the sensitivity achieved from combining the negative surveillance sampling results from multiple sources (Cannon, 2002). The general formula is $Se = 1 - \prod_{l=1}^{L}(1 - Se_l)$, where each of the Se_l value is the confidence level achieved from sampling subpopulation l, with $Se_l = 1 - (1 - P_{T_l})^{n_l} = 1 - (1 - \gamma_l P)^{n_l}$, where n_l is the number of samples from subpopulation l and P is the design prevalence.

3. Estimation of prevalence in the targeted sampling framework

The presentation thus far has focused on demonstrating freedom from disease. The logical extension of targeted sampling is the estimation of prevalence. The estimator of prevalence for a simple random sample of size n is given by

$$\hat{P}_{SRS} = \frac{\sum_{i=1}^{n} \delta_i}{n},$$

where the indicator variable is defined as

$$\delta_i = \begin{cases} 1 & \text{if animal } i \text{ is diseased} \\ 0 & \text{otherwise} \end{cases}.$$

Suppose that the entire sample is allocated to a single subpopulation made up of animals that exhibit characteristic T. If the risk ratio and fraction of the population are

known values, and a random sample of size n_T is drawn from this subpopulation, then the targeted sampling estimator of prevalence is given by

$$\hat{P}_{Targ} = \frac{\sum_{i=1}^{n_T} \delta_i}{\gamma n_T}.$$

Two approaches can be used to show that this estimator is unbiased. The first approach follows methodology typically found in survey sampling (e.g., Sarndal et al., 1992), which yields

$$\begin{aligned} E[\hat{P}_{Targ}] &= E\left[\frac{\sum_{i=1}^{n_T} \delta_i}{\gamma n_T}\right] \\ &= \frac{P}{P_T}\frac{\sum_{i=1}^{n_T} E[\delta_i]}{n_T} \\ &= \frac{P}{P_T}\frac{n_T P_T}{n_T} \\ &= P \end{aligned}$$

This approach shows that $\hat{P}_{Targ}$ satisfies the conditions for design unbiasedness, although the sample design is not measurable (Sarndal et al., 1992, p. 33).

The second approach is model-based and assumes that the number of diseased animals observed in the sample is $X = \sum_{i=1}^{n_T} \delta_i$ and is distributed as a binomial random variable (i.e., $X \sim$ Binomial (n_T, P_T)). The expected value of the estimator is

$$\begin{aligned} E[\hat{P}_{Targ}] &= \frac{E[X]}{\gamma n_T} \\ &= \frac{P}{n_T P_T} E[X] \\ &= \frac{P}{n_T P_T} n_T P_T \\ &= P \end{aligned}$$

Regardless of the method of proof, the estimator is unbiased. However, note that prevalence is assumed to be a known value through γ, and n_T is not necessarily an integer.

Assuming the population is large in relation to the sample size, the variance of the estimator is given by

$$Var[\hat{P}_{Targ}] = \frac{P(1-P)}{\gamma n_T}.$$

The SRS and targeted estimators for variance are approximately equal (i.e., $Var[\hat{P}_{SRS}] = P(1-P)/n \approx P(1-P)/\gamma n_T$). Thus, both the probability of detection and the variance of the targeted sampling estimator will be approximately equal to those derived from the equivalent simple random sample. This result ignores the minor differences in the definition of γ values and the need for the sample sizes to be integer values.

4. Estimation with uncertainty in the risk ratio and population fraction

The estimators for targeted sampling assume that the relationship between the design prevalence and the prevalence in each subpopulation are known values. In practice, however, RR and f_T will seldom (if ever) be known with certainty. Unlike many survey applications, where auxiliary information is either enumerated in the sampling frame or collected as part of a multi-stage or multi-phase sample design, the most common sources of risk ratio estimates are published studies (e.g., Wilesmith et al., 1992; Baylis et al., 2002), while the fraction of the population exhibiting the characteristic may come from an area or national survey that provides demographic information (e.g., NAHMS, 1997, 2003). In the absence of empirical data, uncertainty information is often provided by expert opinion. Regardless of the source of epidemiologic information, probability distributions that describe the risk ratio and fraction of the population exhibiting the characteristic are required.

Suppose the random variables $\widetilde{RR}$ and $\tilde{f}_T$ can be described using an appropriate probability distribution, with possible examples being $\widetilde{RR} \sim Gamma(\mu, \sigma^2)$, $\tilde{f}_T \sim Beta(a, b)$. The distributions of possible point values and subpopulation prevalence values are derived from $\tilde{\gamma} = \frac{\widetilde{RR}}{\tilde{f}_T \widetilde{RR} + (1-\tilde{f}_T)}$, and $\tilde{P}_T = \frac{\widetilde{PRR}}{\tilde{f}_T \widetilde{RR} + (1-\tilde{f}_T)}$.

The distribution describing the point values contains both a product and ratio of random quantities. This introduces two concerns. First, it is unlikely that a convenient closed-form solution will exist to describe the distribution of $\tilde{\gamma}$; the solution that we present later uses Monte Carlo methods to approximate the distribution. The second concern is that the estimator, $\tilde{\gamma}$, will be biased, as is the case with any ratio of random variables that are not linearly related (Mood et al., 1974, p. 181).

4.1. Estimation of prevalence with uncertain point values

A consequence of uncertainty in the point values is a bias in the estimated prevalence. The approximate bias can be derived from a Taylor's series approximation (Mood et al., 1974) and is found as follows

$$\begin{aligned} E[\hat{\tilde{P}}_{Targ}] &= E\left[\frac{\sum_{i=1}^{n_T} \delta_i}{\tilde{\gamma} n_T}\right] \\ &= \frac{1}{n_T} E\left[\frac{\sum_{i=1}^{n_T} \delta_i}{\tilde{\gamma}}\right] \\ &\approx \frac{1}{n_T}\left[\frac{E[\sum_{i=1}^{n_T} \delta_i]}{E[\tilde{\gamma}]} - \frac{Cov[\sum_{i=1}^{n_T} \delta_i, \tilde{\gamma}]}{E[\tilde{\gamma}]^2} + \frac{E[\sum_{i=1}^{n_T} \delta_i]}{E[\tilde{\gamma}]^3} Var[\tilde{\gamma}]\right], \end{aligned}$$

where the number of diseased animals is distributed as $\sum_{i=1}^{n_T} \delta_i \sim Binomial(n_T, P_T)$.The numbers of diseased animals $(\sum_{i=1}^{n_T} \delta_i)$ and the estimated points to apply to the sample are uncorrelated in applications where the risk ratio and subpopulation fractions are estimated from independent sources. Therefore, the expected value reduces to $E[\hat{\tilde{P}}_{Targ}] \approx P + P/E[\tilde{\gamma}]^2 Var[\tilde{\gamma}]$. The bias term $P/(E[\tilde{\gamma}]^2) Var[\tilde{\gamma}]$ is positive in all applications so the population prevalence will tend to be overestimated.

4.2. Adjusting sample size for disease detection with uncertain point values

Uncertainty in the point values affects the determination of the necessary sample size to demonstrate freedom

from disease. Note that in this application the prevalence in the targeted subpopulation is now assumed to be a random quantity that is a function of the design prevalence, $\widetilde{RR}$, and $\tilde{f}_T$. The question then becomes; How must the sample size be adjusted if prevalence in the targeted population is a random variable describing uncertainty about its true value?

This uncertainty requires that sample size be assessed with respect to the possible range of $\tilde{P}_T$. The probability density function will be denoted by $f_{\tilde{P}_T}(\tilde{P}_T)$. To detect disease at the specified design prevalence, a 0–1 random variable X is defined as

$$X \sim Bernoulli(\, p = 1 - (1 - \tilde{P}_T)^{n_T}),$$

which implies that

$$X = \begin{cases} 0 \text{ when no diseased animals are found.} \\ 1 \text{ otherwise.} \end{cases}$$

If the true prevalence in the targeted population were known, then n_T is selected so that $p(\text{detection in a targeted sample of size } n_T) = p(X = 1) = 1 - (1 - P_T)^{n_T} = 1 - \alpha$.

Given uncertainty about $\tilde{P}_T$, the probability of detection is found by using the conditional probability, so

$$p(X = 1) = 1 - \alpha = \int_{-\infty}^{\infty} p(X = 1 | \tilde{P}_T = P_T)\, f_{\tilde{P}_T}(P_T)\, \mathrm{d}\tilde{P}_T$$

$$1 - \alpha = \int_{0}^{\infty} (1 - (1 - P_T)^{n_T})\, f_{\tilde{P}_T}(P_T)\, \mathrm{d}\tilde{P}_T.$$

The appropriate sample size is found by solving the integral for n_T, however, as mentioned previously, it is unlikely that $f_{\tilde{P}_T}$ will have a closed-form solution. A solution is to use Monte Carlo methods to generate instances of the integrand that can be averaged, then the average is solved (using a search algorithm) for n_T (i.e., choose n_T so that $\left(\sum_{j=1}^{J} (1 - (1 - \tilde{P}_{T_j})^{n_T})/J\right) \approx 1 - \alpha$ where $\tilde{P}_{T_j}$ is a random draw from the prevalence distribution and J is the number of samples of the Monte Carlo model for $\tilde{P}_T$).

4.3. Estimating population prevalence from a targeted sample

After n_T is determined and a targeted sample is collected, the process for estimating the population prevalence (and its standard error) remains. In this case, estimation is conditional on $Y = \sum_{i=1}^{n_T} \delta_i$ and the distribution of $\tilde{\gamma}$ values. The expected value of the prevalence estimator is given by

$$\begin{aligned} E[\hat{P}_{\widetilde{Targ}}] &= E_\gamma[E_Y[\hat{P}_{\widetilde{Targ}}]] \\ &= E_\gamma\left[E_Y\left[\frac{Y}{n_T\tilde{\gamma}}\right]\right] = E_\gamma\left[\frac{n_T P_T}{n_T\tilde{\gamma}}\right] \\ &= P_T \times E_\gamma\left[\frac{1}{\tilde{\gamma}}\right] \\ &= P_T \times \int_0^\infty \frac{1}{\tilde{\gamma}}\, f_\gamma(\tilde{\gamma})\, \mathrm{d}\tilde{\gamma} \end{aligned}$$

in the case where uncertainty about $\tilde{\gamma}$ is represented by a continuous distribution. In most practical applications, the distribution of $\tilde{\gamma}$ will be estimated using Monte Carlo methods so that $E_\gamma(1/\tilde{\gamma})$ will simply be the mean of a simulated distribution for $1/\tilde{\gamma}$.

The variance of the population prevalence is given by

$$\begin{aligned} Var[\hat{P}_{\widetilde{Targ}}] &= E_\gamma[Var_Y[\hat{P}_{\widetilde{Targ}}]] + Var_\gamma[E_Y[\hat{P}_{\widetilde{Targ}}]] \\ &= E_\gamma\left[Var_Y\left[\frac{Y}{n_T\tilde{\gamma}}\right]\right] + Var_\gamma\left[E_Y\left[\frac{Y}{n_T\tilde{\gamma}}\right]\right] \\ &= E_\gamma\left[\frac{n_T P_T(1-P_T)}{{n_T}^2\tilde{\gamma}^2}\right] + Var_\gamma\left[\frac{n_T P_T}{n_T\tilde{\gamma}}\right] \\ &= \frac{P_T(1-P_T)}{n_T} E\left[\frac{1}{\tilde{\gamma}^2}\right] + P_T^2 Var\left[\frac{1}{\tilde{\gamma}}\right] \end{aligned}$$

Monte Carlo methods are used to determine the necessary moments of $1/\tilde{\gamma}^2$ and $1/\tilde{\gamma}$.

Substituting the sample-based estimates for P_T and Monte Carlo based approximations for $E[1/\tilde{\gamma}^2]$ and $Var[1/\tilde{\gamma}]$ provides a sample-based variance estimator. The standard error of $\hat{P}_{\widetilde{Targ}}$ is the square root of its variance.

For low-prevalence applications, the Wald confidence interval, given by $(\hat{P} \pm z_{\alpha/2}\sqrt{\text{var}[\hat{P}]})$, has an achieved coverage rate that is generally much lower than the nominal $(1 - \alpha/2)$ value. The score confidence interval (Agresti and Coull, 1998), is recommended for this application, with the boundaries of the confidence interval given by

$$\left[\hat{P} + \frac{z_{\alpha/2}^2}{2n} \pm \left[z_{\alpha/2}\sqrt{[\text{var}[\hat{P}] + z_{\alpha/2}^2/(4n)]/n}\right]\right] / (1 + z_{\alpha/2}^2/n).$$

Nanthakumar and Selvavel (2004) summarize the performance of other confidence interval techniques for low-prevalence applications.

4.4. Quantifying the impact of bias in point values

Information regarding the epidemiology of the disease may come from studies conducted on different populations, or it may be based on expert opinion, particularly when the disease in question is not known to exist in the population. A concern in these situations is the possibility for the resulting estimated distribution for $\tilde{\gamma}$ to be biased (Garthwaite et al., 2005). A solution that is often employed for disease detection applications is to use conservative estimates of the epidemiologic parameters. This reduces the point values assigned to animals exhibiting the characteristic and increases the point value assigned to apparently healthy animals. This approach leads to conservative assumptions about the prevalence in the targeted population and results in a greater level of confidence than the stated $(1 - \alpha)$ level for disease detection applications.

Understanding the effect of bias on the prevalence estimator is more difficult because the effect of bias on the resulting inference needs to be evaluated by considering the magnitude of the bias with respect to the variance of the prevalence estimator. The bias ratio is a useful measure of the performance of a biased estimator (Sarndal et al., 1992). It is given by

$$BR(\hat{P}_{\widetilde{Targ}}) = \left| B(\hat{P}_{\widetilde{Targ}}) / \sqrt{Var(\hat{P}_{\widetilde{Targ}})} \right|,$$

where $B(\hat{P}_{\widetilde{Targ}})$ is bias of the estimator. The bias ratio can be used to estimate the effect of bias on the coverage

Table 1
Parameters used in the construction of the low- and high-prevalence populations.

Population	P	P_T	P_O	f_T	RR	$\widetilde{RR}$	$\tilde{f}_T$
Low prevalence	6.35×10^{-6}	5.0×10^{-5}	5.0×10^{-6}	0.03	10	*Normal* (10,2) where $RR > 0.1$	*Beta* (5, 161.6)
High prevalence	0.0019	0.01	0.001	0.10	10	*Normal* (10,2) where $RR > 0.1$	*Beta* (5,45)

probability of the confidence interval of $\hat{P}_{\widetilde{Targ}}$; where coverage probability measures the fraction of repeated samples whose estimated confidence intervals will overlap the true population prevalence. Under a large sample normality assumption, bias ratios of less than 0.1 indicate that bias is of little practical concern because the achieved coverage of a 95% confidence interval will still be very close to the nominal value (e.g., 94.9% in the case of a 10% bias). Even a bias ratio of 0.5 can be considered acceptable because at this level the coverage probability of a 95% confidence interval may still be greater than 92% (Sarndal et al., 1992).

Suppose surveillance is planned to detect disease with 95% confidence at the design prevalence P. Assume that only two subpopulations exist, all sampled animals will be selected from the subpopulation of animals that exhibit the characteristic T, and ignore the effect of uncertainty in the point value. In this situation, the bias ratio can be approximated by

$$BR(\hat{P}_{Targ}) \approx \frac{\sqrt{3}(\hat{P}_{Targ} - P)}{P},$$

which is just the bias expressed as a proportion multiplied by $\sqrt{3}$. Substituting in the appropriate values of P_T, P and γ provides an approximation to the upper and lower bounds for γ values so that the bias ratio is less than 0.5. These bounds are given by $\gamma_{low,high} = \gamma \frac{2\sqrt{3}}{(2\sqrt{3} \pm 1)} = (0.78\gamma, 1.41\gamma)$

This admittedly crude approximation suggests that a bias on the order of 20–40% in the point value will have only a minimal influence on the estimator of prevalence in a targeted sampling application.

5. Simulation study

A simulation study was conducted to compare targeted sampling (Targ) with simple random sampling without replacement (SRS). The simulation was written in the R language (R, 2005). The simulation study was carried out on two artificial populations, with the first mimicking a surveillance application where the prevalence of the disease was high (i.e., ~1 per 500) while the second population mimics a very low-prevalence application (<1 per 100,000). To facilitate comparisons, the design prevalence for disease detection was set to true disease prevalence in the populations (i.e., p(1 or more diseased animals are found in a sample) = 0.95).

To demonstrate the effect of estimated $\tilde{\gamma}$ and $\tilde{P}_T$ values, the risk ratio was described as $\widetilde{RR} \sim Normal(\mu, \sigma^2)$, where only $\widetilde{RR}$ values greater than 0.1 are considered, while the distribution for the fraction of the population exhibiting the characteristic was $\tilde{f}_T \sim Beta(a, b)$. A summary of the parameters defining each population is given in Table 1 and the distributions for the low prevalence example are given in Fig. 3. These two examples are used to illustrate the effect of the multiple sources of bias in the estimators associated with the various approximations.

Summary statistics for the distribution of $\tilde{\gamma}$ are provided in Table 2. The difference in the points values for prevalence estimation and detection was as large as 2.8% for the high-prevalence population. The discrepancy

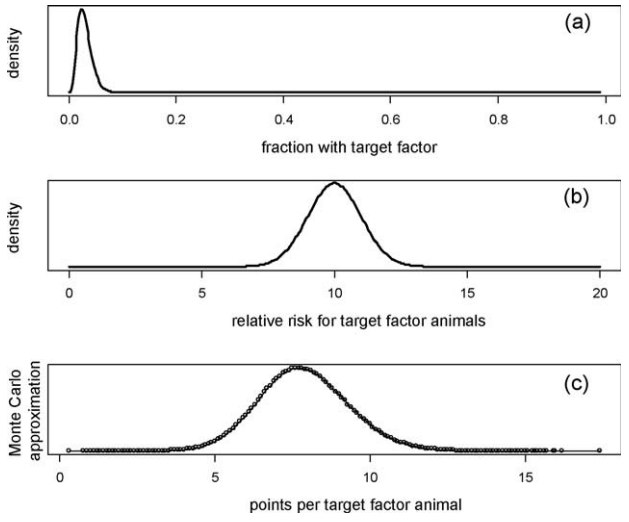

Fig. 3. Monte Carlo based approximations of the distribution describing the uncertainty in the fraction of the population exhibiting the characteristic ($\tilde{f}_T$), risk ratio ($\widetilde{RR}$), and the distribution of points for each targeted sample element (γ).

Table 2
Summary and descriptive statistics for the point values for the two populations.

Population	$\gamma = \ln(1-P_T)/\ln(1-P)$ Detection-based point value	$\gamma = P_T/P$ Prevalence-based point value	Monte Carlo approximated value for $E[\tilde{\gamma}]$	Monte Carlo based standard error $\sqrt{Var[\tilde{\gamma}]}$	Percent bias in $\tilde{\gamma}$ defined by $100(E[\tilde{\gamma}]-\gamma)/\gamma$
Low prevalence	7.8742	7.8740	7.886	1.45	0.16
High prevalence	5.2846	5.2632	5.411	1.48	2.78

Table 3
Metrics for comparing the performance of simple random sampling against the two versions of targeted sampling. The sample sizes are based on the size required to ensure a probability of 0.95 that at least one infected animal would be found in each sample. The low-prevalence data set is representative of surveillance application for extremely rare disease, such as bovine spongiform encephalopathy in the U.S. The high-prevalence population is representative of the OIE standards for surveillance such as bovine brucellosis.

Population	Sampling Method	Sample Size n_*	Confidence in detection *Conf*	Percent Bias of prevalence estimator $B(\hat{P}_*)$	Percent Bias of variance estimator $B(\text{var}[\hat{P}_*])$	Conf. Interval coverage percentage	Bias ratio
Low prevalence	*SRS*	471,767	95.02	0.03	0.01	96.7	<0.00
	Targ	59,914	94.98	−0.04	0.00	96.7	<0.00
	$\widetilde{Targ}$	63,102	95.01	3.48	3.25	95.6	0.05
High Prevalence	*SRS*	1575	95.00	0.02	0.03	96.7	<0.00
	Targ	299	95.06	−0.05	0.92	96.7	<0.00
	$\widetilde{Targ}$	313	95.03	2.23	3.87	95.4	0.04

SRS = simple random sample; *Targ* = targeted sampling with known *RR* and f_T; $\widetilde{Targ}$ = targeted sampling with uncertain epidemiological parameters $\tilde{RR}$ and $\tilde{f}_T$.

between the two definitions of γ is inconsequential for the low-prevalence population.

The sample sizes used in the simulation study were based on a surveillance application where the goal was the detection of disease with an α-level of 0.05. Sample sizes across the two populations and the three sampling methods ranged from 299 to 471,767 (Table 3).

The purpose of the simulation study is to illustrate the potential efficiency of targeted sampling to detect disease and to estimate prevalence. The simulation drew $J = 2{,}000{,}000$ realizations from the population using SRS and targeted sampling. The first metric of interest was the percentage of realizations that contained at least one infected animal, which is given by

$$Conf_* = \frac{\text{Number of realizations with at least one infected animal}}{J},$$

where $* = SRS, Targ, \widetilde{Targ}$ indicates the method of sampling and estimation (i.e., *Targ* assumes *RR* and f_T are known with certainty while $\widetilde{Targ}$ assumes these quantities are uncertain).

For each type of sampling, the appropriate estimators for prevalence and variance of the prevalence estimator were calculated. The mean of the estimates was used to assess the potential relative bias, $B(\cdot)$, of each estimator using

$$B(\hat{P}) = 100 \times \frac{\left(\sum_{j=1}^{J} \hat{P}_j/N\right) - P}{P}.$$

The relative bias of the sample-based variance estimator was assessed using

$$B(\text{var}[\hat{P}]) = 100 \times \frac{\left(\sum_{j=1}^{J} \text{var}[\hat{P}_j]/N\right) - Var[\hat{P}]}{Var[\hat{P}]},$$

where $Var[\hat{P}]$ was the variance of the estimator across the 2,000,000 realizations.

The bias ratio and the achieved confidence interval coverage rate were calculated in order facilitate comparisons between the different sampling methods and estimators.

6. Simulation results

The results of the simulation study are summarized in Table 3. As expected, one or more infected animals were found in approximately 95% of all samples, regardless of the sampling method, though the number of samples needed to achieve this level of confidence was between 5 and 7 times larger when simple random sampling was used. The simulation results illustrate the unbiasedness of the *SRS* and *Targ* estimators of prevalence. On the other hand, the estimator of prevalence, $\hat{P}_{\widetilde{Targ}}$, that accounts for the uncertainty in our knowledge of the epidemiology of the disease, overestimates the true prevalence by between 2.2 (high-prevalence situation) and 3.5% (low-prevalence situation). The bias ratio, $BR(\hat{P}_{\widetilde{Targ}})$, was less than or equal to 0.05 for both populations. This magnitude of the bias ratio indicates that bias will have little effect on the coverage probability of the associated confidence interval. The estimated bias, based on the Taylor's series approximation of the expected value (i.e., $E[\hat{P}_{\widetilde{Targ}}] \approx P + P/(E[\tilde{\gamma}]^2)Var[\tilde{\gamma}]$), was a 2.1 and 4.0% overestimate of the true prevalence, for the high and low-prevalence population, respectively. These values are similar to the percent bias in the simulation study (i.e., a 2.2 and 3.5% over-estimate of true prevalence), which suggests that the assumptions used with the Taylor's series approximation are reasonable and that the targeted sampling estimator will consistently over-estimate the true prevalence. As expected, there was no noticeable bias in the sample-based variance estimators $\text{var}[\hat{P}_{SRS}]$ and $\text{var}[\hat{P}_{Targ}]$, while the bias in $\text{var}[\hat{P}_{\widetilde{Targ}}]$ was 3.25

and.3.9% for the low and high-prevalence populations, respectively.

The achieved coverage rate for the confidence intervals based on the score method (Agresti and Coull, 1998) consistently exceed the nominal 95%, with the coverage rate for the targeted sampling estimator consistently being closest to the nominal value. In contrast, the coverage rates for the Wald-based confidence intervals ranged from only 80–88%, regardless of the population, sampling method, and estimator (results not presented).

7. Conclusions

Targeted sampling is appropriate for surveillance applications where the primary objective is the detection of disease, with the estimation of prevalence being necessary when disease is found. The optimal allocation of samples is to draw all samples from the subpopulation with the highest point value, regardless of whether interest is in detection or estimating prevalence. If the proportion of the population with the characteristic is small and its risk ratio is high, inferences based on targeted sampling will either be very precise in comparison to those derived from a simple random sample of equal size, or a much smaller total sample size will produce estimates of equivalent precision.

The drawbacks associated with targeted sampling are both theoretical and practical. The need to acquire accurate epidemiologic information, particularly for rare diseases, is probably the most challenging aspect of targeted surveillance. The theoretical concern is the complexity associated with the need to incorporate the estimation of point values and complexity of determining appropriate variance estimators and sample sizes. Unlike many survey sampling applications, there are few simple formulas and the solutions may depend on search algorithms and Monte Carlo methods. While none of the estimators are unbiased when the point values are unknown, the biases associated with the approximations are small enough to be of little practical concern.

The reliance on estimated values and models would lead some to argue that targeted surveillance is not an appropriate approach. Never the less, the use of targeted sampling cannot be avoided in applications where the population (or design) prevalence is low and the collection of samples from apparently healthy animals is difficult, impractical, or inappropriate (e.g., the collection of samples from healthy animals when testing for bovine spongiform encephalopathy). Another point to consider is that many surveillance samples are already a form of targeted sampling. For example, surveillance samples collected at slaughter for diseases such as *Brucella abortus* are a targeted sample because breeding cows that abort also have a higher probability of being culled and sent to slaughter. Ignoring the targeted nature of sampling for this application leads to over-estimating prevalence because these samples should be assigned a point value larger than 1.

This study describes animal-level targeted sampling, but the results are also applicable to inferences at the herd- and zone-level using modifications of the results provided by Cannon (2002) for combining surveillance information. Extensions to two-stage surveillance designs are also possible by generalizing the results of Cameron and Baldock (1998). However, the complexity of applications that include both animal- and herd-level risk factors is beyond the scope of this study.

In conclusion, targeted sampling is appropriate for low-prevalence surveillance applications. In these applications, targeted sampling allows for substantial reductions in the required sample size. Nevertheless, it is imperative that uncertainty in the estimated point values be incorporated into the design of targeted surveillance, as well as in the analysis of the resulting sample.

Acknowledgements

Support for this project was provided by the National Surveillance Unit at the Centers for Epidemiology and Animal Health which is part of the Animal Plant Health Inspection Service, U.S. Department of Agriculture.

References

Agresti, A., Coull, B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. Am. Stat. 52, 119–126.
Baylis, M., Goldmann, W., Houston, F., Cairns, D., Chong, A., Ross, A., Smith, A., Hunter, N., McLean, A.R., 2002. Scrapie epidemic in a fully PrP-genotyped sheep flock. J. Gen. Virol. 83, 2907–2914.
Cameron, A.R., Baldock, F.C., 1998. Two-stage sampling in surveys to substantiate freedom from disease. Prev. Vet. Med. 34, 19–30.
Cannon, R.M., 2002. Demonstrating disease freedom-combining confidence levels. Prev. Vet. Med. 52, 227–249.
Christensen, J., Gardner, I.A., 2000. Herd-level interpretation of test results for epidemiologic studies of animal diseases. Prev. Vet. Med. 45, 83–106.
Cochran, W.G., 1977. Sampling Techniques, 3rd Edition. Wiley, New York.
Garthwaite, P.H., Kandane, J.B., O'Hagan, A., 2005. Statistical methods for eliciting probability distributions. J. Am. Stat. Assoc. 100, 680–700.
NAHMS, 1997. Part I: Reference of 1997 Beef Cow-calf Management Practices. USDA:APHIS:VS, CEAH, National Animal Health Monitoring System, Fort Collins, CO #N238.398. June 1997.
NAHMS, 2003. Phase II: Scrapie: Ovine Slaughter Surveillance Study 2002–2003. USDA:APHIS:VS, CEAH, National Animal Health Monitoring System, Fort Collins, CO #N419.0104. January 2004.
Nanthakumar, A., Selvavel, K., 2004. Estimation of proportion of success from a stratified population: a comparative study. Comm. In. Stat. 33, 2245–2257.
Mood, A.M., Graybill, F.A., Boes, D.C., 1974. Introduction to the Theory of Statistics, 3rd ed. McGraw Hill, NY.
OIE. 2006. Surveillance for bovine spongiform encephalopathy. http://www.oie.int/eng/normes/mcode/en_chapitre_3.8.4.htm.
Prattley, D.J., Morris, R.S., Cannon, R.M., Wilesmith, J.W., Stevenson, M.A., 2007a. A model (BSurvE) for estimating the prevalence of bovine spongiform encephalopathy in a national herd. Prev. Vet. Med. 80, 330–343.
Prattley, D.J., Morris, R.S., Cannon, R.M., Wilesmith, J.W., Stevenson, M.A., 2007b. A model (BSurvE) for evaluating national surveillance programs for bovine spongiform encephalopathy. Prev. Vet. Med. 81, 225–235.
R Development Core Team, 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Sarndal, C.E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer-Verlag, New York.
Tavornpanich, S., Gardner, I.A., Carpenter, T.E., Johnson, W.O., Anderson, R.J., 2006. Evaluation of cost-effectiveness of targeted sampling methods for detection of Mycobacterium avium subsp paratuberculosis infection in dairy herds. Am. J. Vet. Res. 67, 821–828.
Wilesmith, J.W., Ryan, J.B., Hueston, W.D., 1992. Bovine spongiform encephalopathy: case-control studies of calf feeding practices and meat and bonemeal inclusion in proprietary concentrates. Res. Vet. Sci. 52, 325–331.